Multiscale Emotion Representation Learning for Affective Image Recognition

Haimin Zhang and Min Xu, *Member, IEEE*

***Abstract*—Recognition of emotions conveyed in images has attracted increasing research attention. Recent studies show that leveraging local affective regions helps to improve the recognition performance. However, these studies do not consider features from the broad context of the local affective regions, which could provide useful information for learning improved emotion representations. In this paper, we present a region-based multiscale network that learns features for the local affective region as well as the broad context for affective image recognition. The proposed network consists of an affective region detection module and a multiscale feature learning module. The class activation mapping method is used to generate pseudo affective regions from a pretrained deep neural network to train the detection module. For the affective region outputted by the detection module, three-scale features are extracted and then encoded by a kernel-based graph attention network for final emotion classification. We show that integrating features from the broad context is effective in improving the recognition performance. We experimentally evaluate the proposed network for both multi-class emotion recognition and binary sentiment classification on different benchmark datasets. The experimental results demonstrate that the proposed network achieves improved or comparable performance as compared to previous state-of-the-art models.**

***Index Terms*—Affective image recognition, multiscale representation learning, deep neural networks.**

I. Introduction

Emotions, which are universal in humans [1], play a considerable role in people's lives. Research shows that human emotions can be evoked by visual stimuli such as images [2], [3]. With the popularity of social media platforms, where people can easily upload and share images, recent years have seen an increasing interest in developing intelligent algorithms for understanding emotions in images. A variety of applications, such as affective image retrieval [4] and opinion mining [5], will benefit from this research.

Recognition of emotions in images is a complicated task. Unlike image semantics, visual emotions usually have high intra-class variations [6]. Images that convey the same emotion can be taken in very different scenes with various objects. The high intra-class variations make it difficult to learn robust emotion representations. Early studies primarily leverage handcrafted features, which are developed for semantic image analysis or designed based on the psychology and art theory, for emotion representation. The low-level features are usually difficult to be interpreted by humans, therefore there remains a large gap between the features and emotions. In recent years, convolutional neural networks (CNNs) have become the dominant approach for various computer vision tasks such as image classification [7], [8] and scene recognition [9]. For affective image recognition, studies have also shown that CNNs exhibit significantly improved performance compared to handcrafted features [10], [11]. State-of-the-art deep neural networks, *e.g.,* ResNets [7] and DenseNets [12], were originally developed for recognizing visual semantics, such as objects, in images. Although these networks can be directly applied for emotion recognition in images, it is critical to develop models that are specifically for the affective image recognition task to further improve the recognition performance.

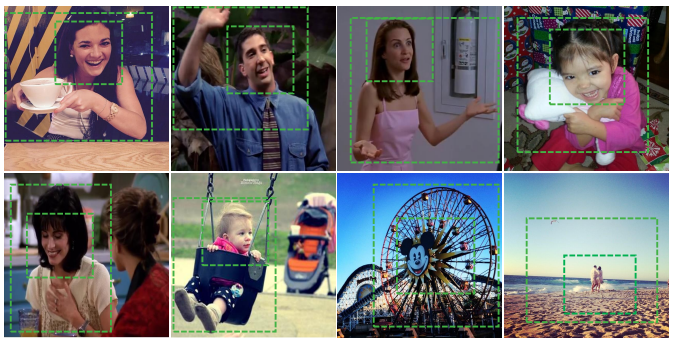

Fig. 1: An illustration that shows emotion clues in an image can be found from multiple scales.

The dominant emotion in an image can usually be found in a local region, while the other parts of the image may reveal a neutral or other non-dominant emotion. Recent studies have shown that leveraging local affective regions helps to improve the recognition performance [11], [13]. A limitation of these studies is that they require emotion region annotations, which may take laborious work to obtain or a computationally intensive procedure to discover. Lee *et al.* [14] considered the human face in an image as the local affective region and proposed a two-stream network that simultaneously extracts features from the face region, detected by an off-the-shelf face detector, and the whole image other than the face region. However, this method can only be applied to tasks in which the images contain human faces. It is not applicable to more general affective image recognition tasks.

As illustrated in Figure 1, emotion clues in an image can usually be found from multiple scales, features extracted from the broad context could be complementary to those extracted from the local affective region for emotion recognition. Multi-

Haimin Zhang and Min Xu (corresponding author) are with the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia (e-mail: Haimin.Zhang@uts.edu.au, Min.Xu@uts.edu.au).

scale representation learning has been studied in different areas including conventional feature design [15], deep representation learning [16], [17] and face detection [18]. It has also been successfully applied for facial expression recognition [19], [20], [21], which suggests that effective emotion features can be extracted from multiscale contexts. However, existing studies have not considered learning multiscale emotion features for the affective image recognition task. In this work, we propose an end-to-end multiscale learning network for this task. The proposed network consists of two modules: an affective region detection module and a multiscale representation learning module. For the affective region outputted by the detection module, we extract multiscale features and then encode the multiscale features with a graph convolutional network for final emotion classification. We show that incorporating features from multiple scales is effective in learning improved emotion representations.

Because most of the existing datasets for image emotion recognition have only image-level annotations. We are unable to directly train our network for detecting local affective regions. Instead, we adopt a weakly supervised learning strategy. We first train a convolutional network for emotion classification using image-level labels and then extract region bounding boxes using the class activation mapping (CAM) method [22]. The obtained bounding boxes are used as pseudo region annotations to train our network for affective region detection. For the affective region detected by the detection module, three-scale features are extracted and then encoded by a kernel-based graph attention network (KGAT), in which the attention weights are computed by similarity comparison in the reproducing kernel Hilbert space (RKHS). Finally, we concatenate the three-scale features for emotion classification. The whole network can be trained in an end-to-end fashion.

The main contributions of this paper can be summarized as follows.

- We propose an end-to-end multiscale learning network that consists of an affective region detection module and a multiscale learning module for recognition of emotions in images. For the detected affective region, three-scale features are extracted and then fused for final emotion classification. We show that our network learns improved emotion representations by integrating features from a broad context.
- A kernel-based graph attention network, in which the attention weights are computed by similarity comparison in the RKHS, is introduced for encoding the three-scale features. We show that our kernel attention method achieves improved performance compared to conventional dot-product attention.
- We experimentally evaluate the proposed network on different benchmark datasets for both emotion recognition and binary sentiment classification. The experimental results demonstrate that our network achieves improved or comparable performance as compared to previous state-of-the-art models.

II. Related Work

A. Affective Image Recognition

Early studies on affective image recognition are mainly conducted on small-scale datasets such as IAPS [23] and ArtPhoto [24]. To model the emotional information in an image, low-level and/or mid-level image features are extracted based on the psychology and art theory. Machajdik *et al.* [23] evaluated a number of low-level features, including color, texture and harmonious composition, for emotion recognition in images. Zhao *et al.* [25] introduced to use features extracted based on the principles of art, including balance, emphasis, harmony, variety, gradation and movement, for emotion representation. Pang *et al.* [4] proposed a method based on the deep Boltzmann machine [26] to learn features from multimodal inputs for emotion classification and cross-modal retrieval. They showed that the learned features are complementary to single-modal features to improve the recognition performance. However, the use of the learned multimodal features alone could not improve the recognition accuracy compared to the use of single-modal features.

Over the past years, researchers have started to leverage deep learning for emotion recognition. Chen *et al.* [27] introduced a visual sentiment classification method using CNNs. They showed that CNN-based methods significantly outperform SVM-based methods for image sentiment annotation and retrieval. In [28], a multitask learning framework was developed for joint emotion classification and emotion distribution regression. This framework helps to tackle the problem with annotating images using hard emotion labels, *i.e.,* each image is labeled with a single emotion category. Rao *et al.* [29] introduced to use multilevel features extracted from a CNN for emotion recognition. This method can extract both low-level features and high-level semantic features. Zhu *et al.* [30] proposed a unified CNN-RNN model in which multilevel features are extracted from a CNN and then fused with a bidirectional recurrent neural network (RNN). Integrating multilevel features, the CNN-RNN model is effective in improving the recognition performance. Pando *et al.* [31] proposed a curriculum-guided strategy to learn emotion features. They showed that the model trained on a large-scale image dataset exhibits impressive generalization performance across datasets.

Recently, researchers have begun to leverage local affective regions for emotion representation. Yang *et al.* [11] proposed a method for affective region localization leveraging an off-the-shelf region proposal tool, *e.g.,* EdgeBoxes [32]. In this framework, thousands of candidate regions are required to be processed for one input image, which is computationally intensive and time consuming. Rao *et al.* [13] proposed a multilevel region-based convolutional neural network based on the feature pyramid network for emotion recognition in images. This network can detect local emotion regions and has been shown to achive improved performance for emotion classification compared to the previous models. However, this network needs to be pretrained on a dataset with region annotations for affective region detection. More recently, Lee *et al.* [14] proposed a two-stream context-aware network that

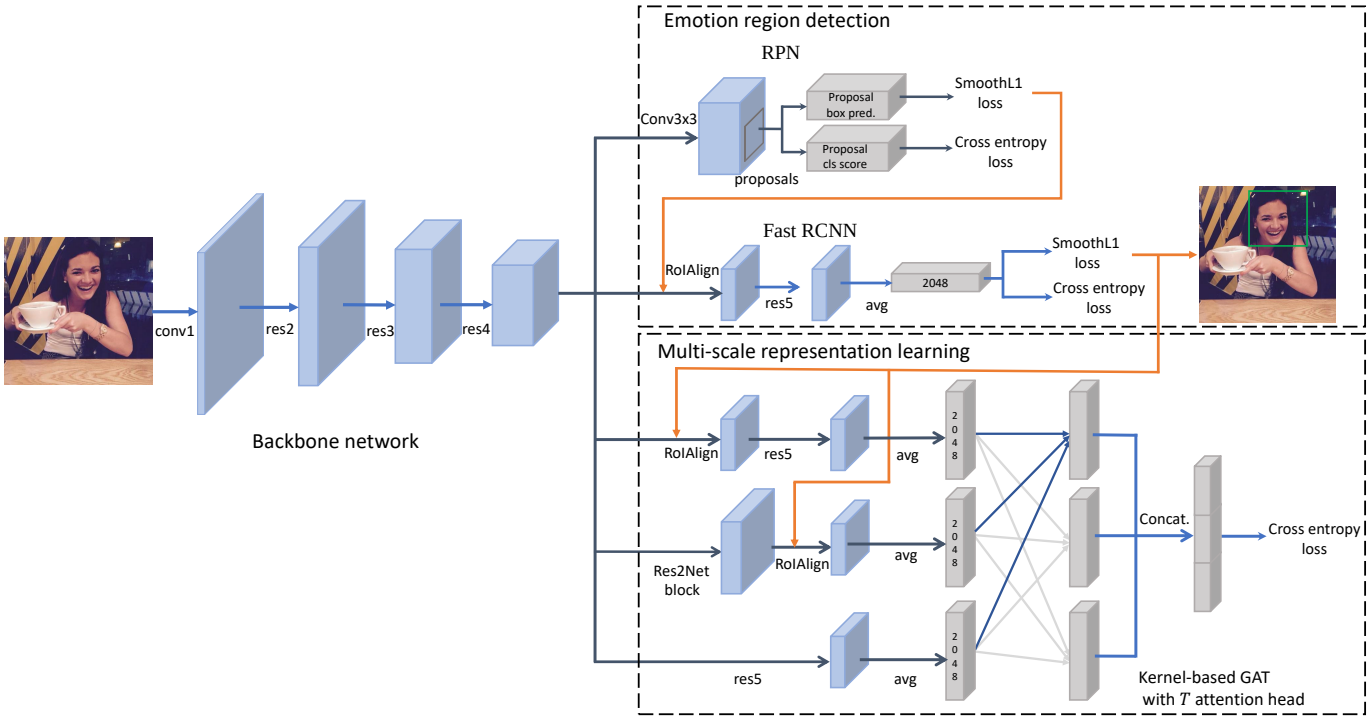

Fig. 2: An overview of the proposed multiscale learning network for emotion recognition in images. This network consists of an affective region detection module and a multiscale feature learning module. The detection module is trained using pseudo affective regions generated with the CAM method. For the detected affective region, three-scale features are extracted and then encoded by a kernel-based graph attention network for final emotion classification.

simultaneously extracts features from the face region and the whole context image. This network can be considered as a two-scale feature learning network; however, it cannot be applied to more general affective image recognition tasks.

B. Graph Convolutional Networks

Graph convolutional networks (GCNs) have been widely applied for learning on graph-structured data such as social networks, citation networks and molecular structures. Existing graph convolutional networks can be categorized into two approaches: the spectral-based approach and the spatial-based approach [33]. The spatial-based approach directly defines convolutions on graphs and updates node features by aggregating local neighborhood information. This is much efficient compared to the spectral-based approach. Velickovic *et al.* [34] introduced graph attention networks (GATs), in which the self-attention mechanism is applied to compute attention weights in aggregation. Zhang *et al.* [35] proposed a method that learns attention weights by similarity comparison in the reproducing kernel Hilbert space using multiple kernels to improve the performance of GATs. Bresson *et al.* [36] proposed residual gated graph convnets (GatedGCNs), integrating edge gates, residual learning [7] and batch normalization [37] into the graph network model. In recent years, researchers have successfully combined GCNs with CNNs for computer vision tasks. Wu *et al.* [38] developed a two-stream GCN-based network for image captioning. Gao *et al.* [39] proposed a tracking framework that uses a spatial-temporal GCN and a context GCN for target appearance modeling and localization.

III. Methodology

Based on the observation that emotion clues in an image can be found from multiple scales, we present a multiscale learning network for recognition of emotions in images. The proposed network consists of an affective region detection module and a multiscale feature learning module. For the affective region detected by the detection module, we extract three-scale features and encode them using a kernel-based graph attention network. Finally, the three-scale features are concatenated for emotion classification. An overview of the proposed network is shown in Figure 2.

Our method involves two steps. First we train a convolutional network and use the obtained model to extract pseudo affective regions with the CAM method. In the second step, we train our multiscale learning network using training images and the pseudo regions. In this section, we first introduce the CAM method for generating pseudo affective regions. Then, we present the details of our multiscale learning network for emotion recognition in images.

A. Pseudo Affective Region Generation

Most of the existing datasets for image emotion recognition have only image-level annotations. We are unable to directly train the affective region detection module because it needs region annotations. Manually annotating affective regions would require laborous work. We instead adopt a weakly supervised

strategy to generate pseudo affective regions and use the pseudo regions to train our detection module for detecting local affective regions.

Specifically, we follow the work of Zhou *et al.* [22] to locate the affective regions using the CAM method. The CAM method leverages a CNN trained using image-level labels and generates class activation maps by projecting back the weights of the output layer onto the feature maps outputted by the last convolutional layer. The class activation map for a category shows the saliency region used by the network to identify that category. It is computed as the weighed sum of the convolutional feature maps. For an image, the value of the class activation map for category c at location (x, y) is defined as follows:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y), \tag{1}$$

where $f_k(x, y)$ denotes the value of the k-th convolutional feature map at spatial location (x, y), and w_k^c denotes the k-th value of the linear transformation weight for predicting class c in the last fully connected layer. For an image, the CAM method can generate a saliency/attention map for each of the network's output categories.

To obtain the bounding box from a class activation map, we adopt the thresholding method. We perform binarization for the class activation map with a threshold value which equals to 20% of the map's maximum value and take the bounding box that covers the largest connected component in the binarized map. This procedure is only performed for class activation maps corresponding to the true labels. The bounding box regions show the most discriminative part of the images that conveys an emotion. After obtaining the bounding boxes, we use them as pseudo region annotations to train our affective region detection subnetwork.

B. The Proposed Multiscale Learning Network

Like the Faster RCNN framework [40], the proposed network for emotion recognition in images is a two-stage architecture. In the first stage, the local affective region is identified. For the identified affective region, a unified multiscale representation is learned for emotion classification in the second stage.

As shown in Figure 2, the proposed network starts with a backbone network, followed by two modules: an affective region detection module and a multiscale feature learning module. For the detection module, we adopt the faster RCNN framework. A region proposal network (RPN) is used to generate candidate regions and a fast RCNN detector is used to determine if a candidate region is an affective region. For the detected affective region, we extract three-scale features and then encode the three-scale features using a kernel-based graph attention network. The three-scale features are finally concatenated together for emotion classification.

We use the ResNet as the backbone network. ResNets adopt shortcut connections which can effectively address the gradient vanishing/exploding problem. The RPN takes the feature maps outputted by the backbone network as input and outputs a set of rectangular region proposals, each with a score indicating the probability of the proposal being an affective region. Specifically, we use a sliding window with a spatial size of 3×3 to traverse the input feature maps. The feature maps within the sliding window are mapped to a lower-dimensional feature space by a shared small network. The obtained feature vector is then fed to two separate fully connected layers: a box classification layer and a box regression layer. At each position the sliding window traverses through, k region proposals, referred to as anchors, located at the sliding window center are simultaneously predicted. Following the work of Ren *et al.* [40], we use $k = 9$ anchors with 3 scales and 3 aspect ratios.

To train the RPN, each anchor is assigned with a binary label indicating whether or not the anchor is an affective region. As in faster RCNN, two kinds of anchors are assigned with a positive label: (1) the anchor with the highest intersection over union (IoU) overlap with the pseudo bounding box; or (2) anchors that have an IoU overlap with the pseudo bounding box higher than 0.7. We assign a negative label to anchors if their IoU ratio with the pseudo bounding box is lower than 0.3. The remaining anchors are not used for the training purpose. The loss function for training the RPN is defined as follows:

$$L_{RPN} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*), \tag{2}$$

where p_i denotes the predicted probability for anchor i being an affective region and p_i^* denotes the label of anchor i. t_i and t_i^* are 4-dimensional vectors denoting the predicted and pseudo bounding box coordinates, respectively. The negative log likelihood function is used as the classification loss L_{cls}, and the robust loss function (smooth L_1) [41] is used as the regression loss L_{reg}. Only positive anchors are used to train the regression loss. The two terms are weighted by $\frac{1}{N_{cls}}$ and $\frac{1}{N_{reg}}$, λ is a hyperparameter.

The Fast RCNN is used to classify affective regions from non-affective regions. It first extracts a feature from the candidate box region using a region of interest align (ROIAlign) layer [42] and then performs classification and regression for the extracted feature. The ROIAlign layer converts the feature maps within a region of interest into small fixed-size feature maps (*e.g.*, 7×7). Unlike region of interest (RoI) pooling [41], which may introduce misalignments between the ROI and the extracted feature maps due to the quantization operation, RoIAlign uses the bilinear interpolation operation [43] to compute feature values at four regularly sampled locations in each RoI bin and then aggregates the obtained results with the max pooling method. RoIAlign can properly align the RoI with the grid used for pooling; thus it extracts more robust features than RoI pooling. The aggregated feature maps are fed to a residual block, which consists of two bottleneck layer, followed by a global average pooling layer. Finally, the obtained feature is taken as input to two separate fully connected layers for classification and bounding box regression. The Fast RCNN detector is used only for affective region detection; it is not used for fine-grained emotion classification at this stage.

The affective region predicted by the detection module

covers the most discriminative part of the input image that conveys an emotion. For the detected affective region, we extract three-scale features for final emotion classification. At the first scale, we extract features from the feature maps produced by the backbone network using a RoIAlign layer. The obtained feature maps are fed to a residual block, followed by a global average pooling layer. We denote the feature vector obtained from this scale by $\boldsymbol{f}_1$. At the second scale, the feature is extracted from a broad context. This is achieved by expanding the receptive field size. Specifically, we utilize two Res2Net bottleneck layers [17] without spatial down sampling. Unlike ResNet bottleneck layers, a Res2Net bottleneck layer utilizes a group of 3×3 convolutions, which are sequentially applied to the input and the output of previous 3×3 convolutions. The Res2Net layer can extract features from a broad context compared to the ResNet bottleneck layer. The RoIAlign layer is used to extract features from the output of the last Res2Net layer. The obtained feature maps are taken as input to an average pooling layer that produces a feature vector $\boldsymbol{f}_2$. At the third scale, the feature maps produced by the backbone network are directly fed to the last residual block of the ResNet, followed by an average pooling layer to generate a feature vector $\boldsymbol{f}_3$.

We introduce a kernel-based graph attention network to encode the three-scale features. Each feature scale is modeled by a graph node. We consider the graph to be complete and self-connected, *i.e.,* each pair of nodes are connected by an edge (undirected). Thus, the graph can be denoted $G = (V, E)$, where $V = \{v_1, v_2, v_3\}$ and E are the node set and the edge set, respectively. In our graph attention layer, a shared linear transformation, which is parameterized by $W \in \mathbb{F}^{D' \times D}$ is applied to all node features. Then, we update each node feature by aggregating neighbourhood features. In aggregation, each neighbour is associated with an attention weight which is computed by a kernel function. For node v_i, the attention weight for its neighbour v_j is defined as follows:

$$\alpha_{ij} = \frac{exp(K(\boldsymbol{f}_i, \boldsymbol{f}_j))}{\sum_{p \in \mathcal{N}(i)} exp(K(\boldsymbol{f}_i, \boldsymbol{f}_p))}, \tag{3}$$

where K denotes a kernel function and $\mathcal{N}(i) = \{j : (v_i, v_j) \in E\}$ denotes the set of v_i's neighbours. The radial basis function (RBF), *i.e.,* $K(\boldsymbol{x}, \boldsymbol{x'}) = exp(-\gamma \parallel \boldsymbol{x} - \boldsymbol{x'} \parallel^2)$ with a hyperparameter γ, is used as the kernel function by default. If we use the linear function as the kernel function, Equation (3) is identical to dot-product attention [44]. In contrast to dot-product attention, we compute the attention weight by comparing the similarity between two node features in the RKHS. By mapping features to the RKHS, our method can achieve improved performance compared to dot-product attention. After obtaining all attention weights, we update the feature for node v_i as follows:

$$\boldsymbol{h}'_i = \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W \boldsymbol{h}_j \right), \tag{4}$$

where σ denotes a nonlinear activation function, *e.g.,* the rectified linear activation unit (ReLU) function.

Instead of using a single attention head, we adopt the multihead attention approach. The kernel-based attention is applied T times, and the outputs of the T attention heads are concatenated as the final node feature:

$$\boldsymbol{h}'_i = \Big\Vert_{t=1}^{T} \sigma \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^t W^t \boldsymbol{h}_j \right), \tag{5}$$

where α_{ij}^t denotes the attention weight for the t-th attention head, $\Vert$ denotes the concatenation operation and W^t is the linear transformation weight matrix in the t-th attention head. With the multihead attention approach, the model can simultaneously extract expressive information from different representation subspaces [44].

The three-scale features are concatenated together and then fed to a fully connected layer for final emotion classification. The affective region detection module and the multiscale representation learning module can be trained in an end-to-end fashion. We use the stochastic gradient descent (SGD) method to train the whole network.

IV. Experiments

We conduct extensive experiments on both multiclass emotion recognition and binary sentiment classification to validate our network. In this section, we first introduce the datasets and implementation details and then present the experimental results and comparisons with previous methods.

A. Datasets

CAER-S [14]. The CAER-S dataset was collected from 79 TV shows. It contains approximately 70,000 images categorized into seven categories, *i.e.,* anger, disgust, fear, happy, sad, surprise and neutral. The images were annotated by three independent annotators. This dataset is split into 70%, 10%, 20% for training, validation and testing, respectively. For fair comparison, we use the same train/validation/test split as the work in Lee *et al.* [14].

FI-8 [10]. The FI-8 dataset was collected from Flickr and Instagram. There are 23,308 images labeled with eight emotion categories defined according to the psychological study of Mikels [45] by Amazon Mechanical Turk workers. Because some images no longer exist on the Internet, only 23,164 images were crawled from the Internet. Following the work in [30], [13], this dataset is split into 80%, 5% and 15% for training, validation and testing, respectively.

IAPSsubset, Abstract, ArtPhoto and EmotionROI [24], [46]. The four datasets are small-scale datasets containing 395, 228, 806, and 1980 images, respectively. We use the four datasets to validate our network for image sentiment classification. Following the experimental setup in [13], we use the 5-fold cross-validation method and report the average of the five validation accuracies as the model performance.

B. Implementation Details

The proposed network is implemented in Pytorch [47]. To generate pseudo affective regions, we first train a ResNet using

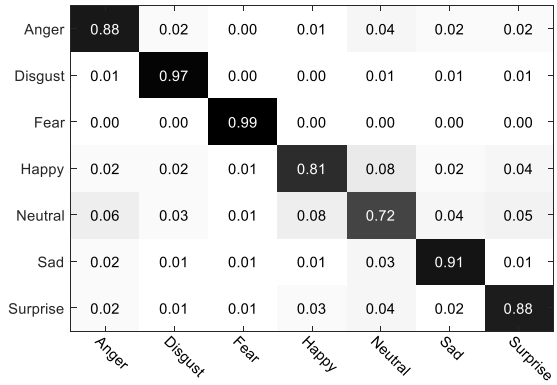

Fig. 3: Confusion matrix for the experiment on CAER-S.

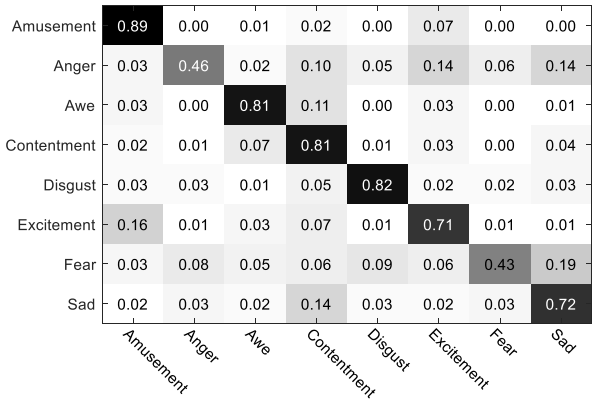

Fig. 4: Confusion matrix for the experiment on FI-8.

Model	Recognition accuracy (%)
AlexNet (off-the-shelf)	47.36
VGGNet (off-the-shelf)	49.89
ResNet-101 (off-the-shelf)	57.33
AlexNet (fine-tuned)	61.73
VGGNet (fine-tuned)	64.85
ResNet-101 (fine-tuned)	68.46
Rao *et al.* [13]	78.35
CAER-Net [14]	73.51
Ours + ResNet-50	**87.47**

TABLE I: Performance of the proposed network on CAER-S and comparison with previous methods.

Model	Recognition accuracy (%)
Sentibank [11]	44.49
Zhao *et al.* [25]	46.52
DeepSentiBank [27]	53.16
PCNN [50]	56.16
AlexNet [51]	58.30
ResNet-101	66.16
MldrNet [29]	67.24
WSCNet [52]	70.07
Zhu *et al.* [30]	73.03
Rao *et al.* + ResNet-101 [13]	75.46
Zhang *et al.* + ResNet-50 [53]	74.84
Zhang *et al.* + ResNet-101 [53]	75.91
Ours + ResNet-50	**76.05**
Ours + ResNet-101	**76.50**

TABLE II: Performance of the proposed network on FI-8 and comparison with previous methods.

image-level labels. The ResNet is initialized using the weights pretrained on ImageNet [48]. We train the ResNet using SGD for 90 epochs with a batch size of 128. The values of weight decay and momentum are set to 0.001 and 0.9, respectively. The learning rate is initialized to 0.001 and divided by 10 at epoch 30 and 60. After the training procedure is finished, we apply the CAM method introduced in Section III-A to the obtained model to generate pseudo affective regions.

For training the proposed multiscale learning network, we use the weights of the pretrained ResNet model to initialize the backbone network. The weights of the remaining layers are initialized by sampling from a zero-mean Gaussian distribution with a standard deviation of 0.01. We follow the strategy used in [40] to train the RPN. The non-maximum suppression (NMS) method is applied to the region proposals based on the classification scores to reduce proposal redundancy. The IoU threshold value for NMS is set to 0.7, and the 2000 top ranked proposals after NMS are selected to train the Fast RCNN. The value of λ is set to 1. The dimension of the hidden features in the graph convolutional layer is set to 256. We use a single layer kernel-based graph attention network with $T = 4$ attention heads. The dropout method [49] with the probability equal to 0.6 is applied to the inputs to the graph network.

The whole network is trained using SGD for 90 epochs with a batch size of 16. The initial learning rate is initialized to 0.0001 for the backbone network and 0.001 for the remaining layers, and reduced by a factor of 10 at epoch 30 and 60. The values of weight decay and momentum are set to 0.0005 and 0.9, respectively. Each channel of input data are normalized to have a zero mean and unit variance. We report the average classification accuracy of three runs as the model performance.

C. Results on CAER-S

We use AlexNet, VGGNet, ResNet-101, the work of Rao *et al.* [13] and CAER-Net [14] as the baseline methods. The recognition accuracy of the proposed network and comparison with the baselines are reported in Table I. We obtain an overall recognition accuracy of 87.47%, significantly outperforming the baseline methods. Compared with the off-the-shelf features extracted from AlexNet, VGGNet and ResNet-101 pretrained on ImageNet, the proposed network improves the performance over 30.0%. The proposed network also achieves at least 19.01% performance improvement compared with fine-tuned AlexNet, VGGNet and ResNet-101. The experimental results demonstrate the superiority of the proposed network over conventional deep neural networks, which are developed for image classification, for the affective image recognition task. CAER-Net [14] is a dual-stream architecture that simultaneously extracts features from the face region and the whole image other the face region, and uses an adaptive fusion network to encode the two-stream features for final emotion classification. The proposed network outperforms the CAER-Net by 13.96%. This shows the advantage of the proposed network over the face region-based method for emotion recognition. To the best of our knowledge, our network achieves the state-of-the-art

Feature scales	Recognition accuracy (%) CAER-S	FI-8
Scale 1	81.55	71.88
Scale 2	82.17	72.33
Scale 3	73.19	67.06
Scale 1 + 2	86.59	75.36
Scale 2 + 3	84.75	74.16
Scale 1 + 2 + 3	87.47	76.05

TABLE III: Ablation: Effect of features from different scales on the overall performance. The ResNet-50 is used as the backbone network for the experiments.

performance on this dataset. Figure 3 shows the confusion matrix for our experiment on this dataset.

D. Results on FI-8

The experimental results on FI-8 and comparison with previous methods are reported in Table II. The proposed network achieves 76.50% overall recognition accuracy using the ResNet-101 as the backbone network. Once again, our network achieves the state-of-the-art performance. The proposed network shows significantly improved performance compared to SentiBank [11], DeepSentiBank [27] and PCNN [50]. It also outperforms AlexNet and ResNet-101 by a large margin. Compared with MldrNet [29] and WSCNet [52], the proposed network improves the recognition accuracy 9.26% and 6.43%, respectively.

Zhu *et al.* [30] proposed a unified CNN-RNN model that fuses features from different levels of a convolutional neural network using a recurrent neural network. Compared with Zhu *et al.*'s method, our network achieves a 3.47% performance improvement. This shows the advantage of our multiscale learning network over the multilevel-based method. The proposed network achieves 1.04% higher recognition accuracy than Rao *et al.*'s [13] multilevel region-based convolutional neural network. Rao *et al.*'s network must be first trained on a dataset that has region annotations for detecting emotion regions. However, most existing affective image recognition datasets have only image-level annotations. In their method, the model is first pretrained on EmotionROI, which contains region annotations, and then applied to other datasets for emotion region detection. However, due to the domain shift across datasets, the model pretrained on EmotionROI might not generalize well to other datasets. In contrast to their method, we use the CAM method to generate pseudo affective regions to train our detection module, therefore our network performs well for affective region prediction for different datasets. Our network outperforms Zhang *et al.*'s [53] by 0.59%. In Zhang *et al.*'s work, the CAM method is integrated into the neural network for emotion recognition in images. Unlike Zhang's work, we first train a ResNet using only image-level labels and extract pseudo affective regions with the CAM method using the pretrained model. This is a preliminary step of our method. In our training stage, the training images and generated pseudo affective regions are used for training our network. At test time, the detection subnetwork first outputs an affective region, then three-scale features are extracted and then fused for final emotion recognition. Figure 4 shows the confusion matrix on the FI-8 dataset.

Model	Recognition accuracy (%) CAER-S	FI-8
Our model w/o KGAT	83.20	73.87
Our model + KGAT	87.47	76.05

TABLE IV: Ablation: Effect of the kernel graph attention network on the overall performance. For 'w/o KGAT', the KGAT module is not used in our model, the three-scale features are directly concatenated for final emotion classification. The ResNet-50 is used as the backbone network for the experiments.

Value of γ'	Recognition accuracy (%) CAER-S	FI-8
Liner kernel (scaled)	85.34	75.16
1/4	87.13	75.85
1/2	**87.47**	76.02
1	87.06	**76.05**
2	86.84	75.63

TABLE V: Impact of the value γ' in the RBF kernel on the overall performance. The ResNet-50 is used as the backbone network for the experiments.

Num. of attention heads	Recognition accuracy (%) CAER-S	FI-8
1	86.71	75.57
2	86.96	75.83
3	87.39	76.04
4	87.47	76.05
5	87.45	76.05
6	87.45	76.03

TABLE VI: Impact of the number of attention heads on the overall performance. The ResNet-50 is used as the backbone network for the experiments.

E. Analysis and Ablation Studies

We conduct an ablation of our model to show the effect of features from different scales on the overall performance. The experiments are conducted on CAER-S and FI-8. Table III reports the experimental results. It can be seen that the first scale feature and the second scale feature show significantly improved performance compared to the third scale feature, demonstrating that locating local affective regions is effective in improving the recognition performance. The second scale feature performs well as compared to the first scale feature, indicating that incorporating features from the broad context is helpful for performance improvement. The joint use of the first scale and second scale features achieves better performance than the use of a single scale feature. The recognition performance is further improved when integrating the third scale features. This demonstrates the effectiveness of our multiscale learning network for emotion recognition in images.

We also conduct an ablation study to show the importance of the kernel-based graph attention network. The experimental results on CAER-S and FI-8 are shown in Table IV. We

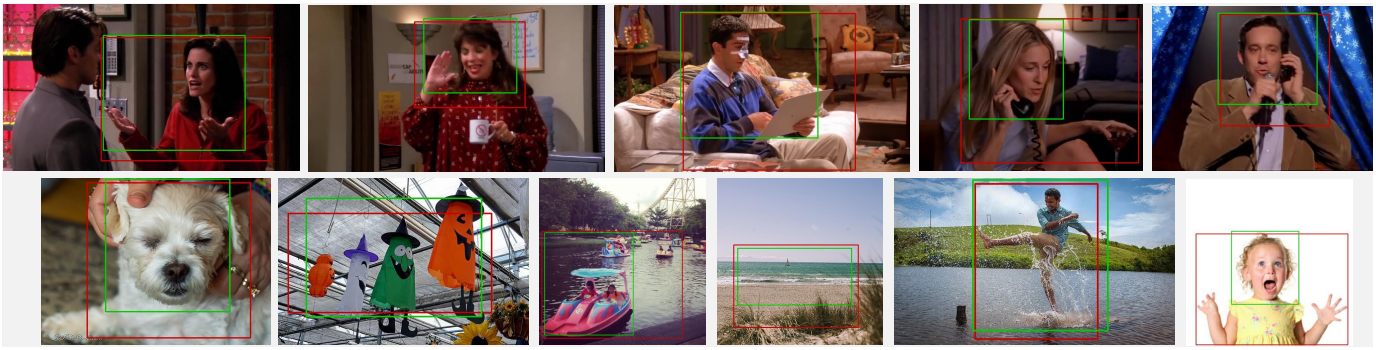

Fig. 5: Samples of affective regions predicted by the proposed network (green rectangles) and the pseudo affective regions (red rectangles) generated with the CAM method. We show that the proposed network performs well in locating local affective regions.

Method	FI-8 (binary sent.)	IAPSsubset	Abstract	ArtPhoto	EmotionROI
DeepSentibank [45]	64.39	85.63	71.19	68.73	70.11
PCNN [50]	75.34	88.84	70.84	70.96	73.58
AlexNet [51]	72.43	84.58	65.49	69.27	71.60
VGGNet-16 [54]	83.05	88.51	72.48	70.09	77.02
ResNet-50 [7]	85.43	89.95	73.07	70.93	79.28
ResNet-101 [7]	85.92	90.13	73.36	71.08	79.67
Curriculum Learning [31]	84.81	–	–	–	–
Zhu *et al.* [30]	84.26	91.38	73.88	75.50	80.52
Yang *et al.* [11]	86.35	92.39	76.03	74.80	81.26
Rao *et al.* + ResNet-101 [13]	87.51	93.66	77.77	77.28	82.94
Zhang *et al.* + ResNet-50 [53]	90.58	95.61	82.59	78.72	84.59
Zhang *et al.* + ResNet-101 [53]	90.97	**95.83**	83.02	79.24	85.10
Ours + ResNet-50	**90.88**	95.47	**82.64**	**78.96**	**84.75**
Ours + ResNet-101	**91.16**	95.58	**83.33**	**79.72**	**85.60**

TABLE VII: Performance of our network for image sentiment (positive or negative) on FI-8, IAPSsubset, Abstract, ArtPhoto and EmotionROI, and comparison with previous methods.

see that integrating the kernel-based graph attention module yields 4.27% and 2.18% performance improvements on the two datasets, respectively. The results demonstrate that our the kernel-based graph attention is effective in improving the overall performance.

Moreover, we study the impact of the parameter γ in the RBF kernel on the overal performance. We rewrite the RBF kernel as $K(\boldsymbol{x}, \boldsymbol{x}') = exp(-\gamma' \parallel \boldsymbol{x} - \boldsymbol{x}' \parallel^2 /d)$, where d denotes the dimension of $\boldsymbol{x}$ and $\boldsymbol{x}'$. The experimental results on CAER-S and FI-8 are shown in Table V. It can be seen that the best performace is achieved when γ' is set to $1/2$ and 1 on the two datasets, respectively. As compared to the linear kernel, *i.e.,* scaled dot-product attention, the RBF kernel improves the performance 2.13% and 0.89% on the two datasets, respectively. This demonstrates the advantage of our kernel-based attention over the dot-product attention.

To show the effects of the multi-head attention approach, we compare the performance of using difference number of attention heads in the KGAT subnetwork. The comparison results are shown in Table VI. We see that using multiple attention heads helps to improve the recognition performance compared to using a single attention head. We also see that using 4 attention heads achieves 0.76% and 0.48% performance gains on the two datasets, respectively, compared with using a single attention head. The results demonstrate the effectiveness of the multi-head attention approach for performance improvement.

Fig. 6: Examples of affective regions detected by our network and attention maps predicted by Zhang *et al.*'s method [53].

Our method is a region-based method for affective image recognition. In the training stage, the regions generated by the CAM method are used for training the affective region detection subnetwork. For the affective region outputted by the detection subnetwork, multiscale features are extracted and then fused for final emotion recognition. Therefore, the performance for predicting local affective regions is significant for the final emotion recognition. Figure 5 demonstrates samples of detected affective regions. We see that these detected affective regions have a high overlap with the pseudo regions generated by the CAM method. This demonstrates our network performs well for locating local affective regions.

Figure 6 compares our network with Zhang *et al.*' model [53] for locating local affective regions. Zhang et al.s model [53] predicts an emotion intensity map for an image, the value of the intensity map at a spatial location indicates the importance of that location revealing an emotion. Both the work can highlight the most important affective region for emotion recognition. Importantly, our network is able to extract features from the broad context of the local affective region, which is usually helpful for learning improved emotion representations.

In many applications with graph convolutional networks, the nodes in a graph have a small number of neighbours, *e.g.*, 1 to 3. For example, in three popular datasets (Cora, Citeseer, Pubmed) [55], there are 2708, 3327, 19717 nodes and 5429, 4732, 44338 edges respectively. The average number of edges per node is 2.005, 1.42 and 2.25 respectively on the three datasets. Graph convolutional networks, such as graph attention networks (GATs), have shown excellent performances for node classification on the three datasets. In our model, the three scale features are modeled with a three node graph. The attention weights in our model are computed by similarity comparison in the reproducing kernel Hilbert space. We show that applying our graph attention network improves the recognition performance. Figure 7 visualizes attention scores computed by comparing the first scale feature with the three scale features on several examples.

F. Image Sentiment Classification

We further validate the proposed network for binary image sentiment classification. The goal of sentiment classification is to classify an image as having a positive sentiment or negative sentiment, that is the general attitude or opinion revealed in the image. The experiments are conducted on FI-8, IAPSsubset, Abstract, ArtPhoto and EmotionROI. For FI-8, we convert the original labels to positive or negative. To reduce overfitting on IAPSsubset, Abstract, ArtPhoto and EmotionROI, we use the weights pretrained on FI-8 to initialize the model before optimization.

The experimental results and comparison with the latest methods are shown in Table VII. We see that the proposed network shows significantly improved performance compared to DeepSentiBank [45] and PCNN [50], as well as AlexNet, VGGNet and ResNet which are state-of-the-art networks for image classification. Zhang *et al.*'s method [53] achieves the best performance among the five most recent methods that were developed specifically for sentiment classification. For using the ResNet-101 as the backbone network, the proposed network improves the performance 0.19%, 0.31%, 0.48% and 0.50% on FI-8, Abstract, ArtPhoto and EmotionROI, respectively, as compared to Zhang *et al.*'s method. Our network also achieves comparable performance as Zhang *et al.*'s method on IAPSubset. The results demonstrate the effectiveness of the proposed network for sentiment classification. In Zhang *et al.*'s work [53], the feature maps highlighted by the predicted CAM map and the original feature maps are fused together for final emotion recognition. Unlike Zhang *et al.*'s work, our method first predicts an affective region. For the affective region, three-scale features are extracted and then fused with a kernel-based graph attention network for final emotion classification. Compared with Zhang *et al.*'s work, our method integrates more context information of the local affective region for emotion representation. The experimental results show that this is effective in improving the recognition performance.

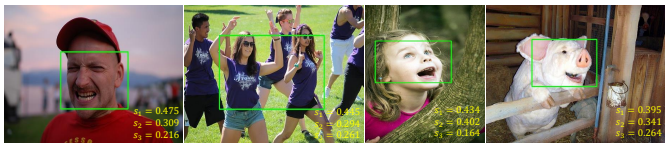

Fig. 7: Illustration of attention scores computed by comparing the first scale feature with the three scale features.

V. Conclusion

In this paper, we presented an end-to-end multiscale learning network for recognition of emotions in images. Our method is inspired by the observation that emotion clues in an image can usually be found from multiple scales. The proposed network is a two-stage architecture. In the first stage, the local affective region is identified. A unified multiscale feature is learned for emotion classification in the second stage. The CAM method is used to generate pseudo affective regions to train the proposed network for affective region detection. Integrating features from the broad context, the proposed network learns improved emotion representations. In addition, we intorduce a kernel-based graph attention network, in which the attention weights are computed by similarity comparison in the RKHS, to encode the features from different scales. We showed that our kernel-based attention is effective in improving the recognition performance compared to conventional dot-product attention. The proposed network was evaluated for multiclass emotion recognition and binary sentiment classification on different benchmark datasets. The experimental results demonstrate that our network achieves improved or comparable performance as compared to previous state-of-the-art methods.

VI. Acknowledgement

We would like to thank the reviewers for reviewing our manuscript.

Appendix A

The RoIAlign layer [42] is used to extract a set of small feature maps for each region of interest. It first divides the region of interest into spatial bins (*e.g.,* 7×7). Then the bilinear interpolation is applied to compute features at four regularly sampled locations in each region of interest bin from the input feature maps, and the obtained features in each bin are aggregated with a pooling function (*e.g.,* average pooling).

Figure 8 shows an illustration of the ROIAlign layer. The dashed grid represents a convolutional feature map, and the solid lines represent 2×2 RoI bins. The blue dots in each bin are the 4 sampled points. The RoIAlign layer computes the value at each sampling point from the nearest grid points in the convolutional feature map using bilinear interpolation.

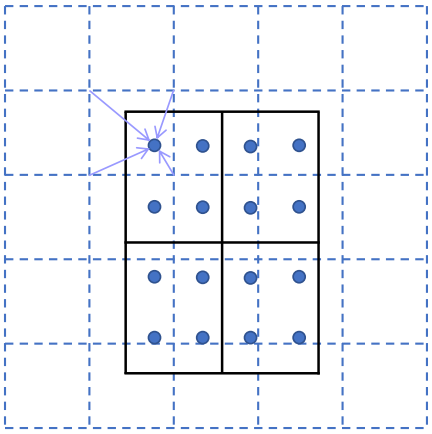

Fig. 8: An illustration of the ROIAlign layer [42].

REFERENCES

[1] P. Ekman, "Emotion in the human face (2e éd.)," 1982.

[2] P. J. Lang, "A bio-informational theory of emotional imagery," *Psychophysiology*, vol. 16, no. 6, pp. 495–512, 1979.

[3] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology," *Biological psychiatry*, vol. 44, no. 12, pp. 1248–1263, 1998.

[4] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[5] Q.-T. Truong and H. W. Lauw, "Visual sentiment analysis for review images with item-oriented and user-oriented cnn," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1274–1282.

[6] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2824–2835, 2018.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[9] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: objects, scales and dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 571–579.

[10] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[11] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018.

[12] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[13] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.

[14] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 143–10 152.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[17] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[18] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[19] Z. Li, S. Wu, and G. Xiao, "Facial expression recognition by multi-scale cnn with regularized center loss," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3384–3389.

[20] H. Li, L. Chen, D. Huang, Y. Wang, and J.-M. Morvan, "3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 2577–2580.

[21] N. Perveen, D. Roy, and K. M. Chalavadi, "Facial expression recognition in videos using dynamic kernels," *IEEE Transactions on Image Processing*, vol. 29, pp. 8316–8325, 2020.

[22] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." *CVPR*, 2016.

[23] P. J. Lang, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," *Technical report*, 2005.

[24] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92.

[25] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 47–56.

[26] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial intelligence and statistics*, 2009, pp. 448–455.

[27] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[28] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network." in *IJCAI*, 2017, pp. 3266–3272.

[29] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, pp. 1–19, 2016.

[30] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition." in *IJCAI*, 2017, pp. 3595–3601.

[31] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[32] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[33] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[34] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," *International Conference on Learning Representations*, 2018, accepted as poster. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[35] H. Zhang and M. Xu, "Graph neural networks with multiple kernel ensemble attention," *Knowledge-Based Systems*, vol. 229, p. 107299, 2021.

[36] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[38] L. Wu, M. Xu, L. Sang, T. Yao, and T. Mei, "Noise augmented double-stream graph convolutional networks for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[39] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4649–4659.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[41] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[43] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[45] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, no. 4, pp. 626–630, 2005.
[46] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 614–618.
[47] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch," *Computer software. Vers. 0.3*, vol. 1, 2017.
[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[50] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[52] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, 2019.
[53] H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Transactions on Multimedia*, 2020.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[55] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.

Haimin Zhang is currently a Postdoctoral Research Fellow in the School of Electrical and Data Engineering, University of Technology Sydney. He received the Bachelor's degree from Zhejiang Sci-Tech University, Hangzhou, China, the Master's degree from Nankai University, Tianjin, China, and the Ph.D. degree from University of Technology Sydney, Ultimo, NSW, Australia. His research interests include multimedia data analytics, computer vision, pattern recognition and machine learning. He serves as a reviewer for several journals including IEEE Transactions on Multimedia and Pattern Recognition.

Min Xu (M'10) is currently an Associate Professor at University of Technology Sydney. She received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2000, the M.S. degree from National University of Singapore, Singapore, in 2004, and the Ph.D. degree from University of Newcastle, Callaghan NSW, Australia, in 2010.

Her research interests include multimedia data analytics, computer vision and machine learning. She has published over 100 research papers in high quality international journals and conferences. She has been invited to be a member of the program committee for many international top conferences, including ACM Multimedia Conference and reviewers for various highly-rated international journals, such as IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology and much more. She is an Associate Editor of Journal of Neurocomputing.